GEENIENNUSTEIDEN LAATUMITTAREISTA JA NIIDEN SOVELTAMISESTA MANSIKAN GENOMIIN

Antti Tuominiemi
Maisterintutkielma
Helsingin Yliopisto
Biotekniikka (MAAT)
Marraskuu 2020

TIIVISTELMÄ

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Osasto — Sektion — Department
Maatalous-metsätieteellinen tiedekunta	Maataloustieteiden osasto

Tekijä — Författare — Author

Antti Tuominiemi

Työn nimi — Arbetets titel — Title

Geeniennusteiden laatumittareista ja niiden soveltamisesta mansikan genomiin

Oppiaine —Läroämne — Subject

Biotekniikka

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Maisterintutkielma	1.11. 2020	37

Tiivistelmä – Referent – Abstract

Eliöiden perimän tutkimuksessa käytettävät sekvensointimenetelmät ovat halventuneet aikaisemmasta, jonka takia käytettävissä olevan genomidatan määrä on kasvanut merkittävästi. DNA:n emäsjärjestyksen tietäminen ei auta eliön toiminnan tutkimisessa ennen kuin se annotoidaan, joka tarkoittaa geenien sijaintien etsimistä ja niiden tuotteiden määrittelyä. Annotointiin käytettävät ohjelmat tekevät virheitä ja niiden tuloksia täytyy arvioida erinäisin keinoin. Käytössä olevan datan määrä osaltaan kannustaa tuottamaan uusia annotaatioita nopeammin ja tämä voi lisätä ihmisperäisten virheiden määrää. Osa ohjelmista käyttää geenitietokantoja, joten niiden sisältämien virheellisten geenien määrä voi kasvaa tulevaisuudessa, jos annotaatioiden laadunarviointi keinoja ei kehitetä paremmiksi.

Tässä tutkimuksessa tarkastellaan korrelaatiota valittujen laatumittarien ja annotaatioiden laadun välillä. Käytettävät laatumittarit voidaan jakaa kahden tyyppisiin, toiset perustuvat geenien perusrakenteisiin ja toiset geenin tuottaman proteiinin vertaamiseen proteiinitietokantaa vastaan. Tutkimuksessa oletetaan, että vertaus referenssiin on luotettava keino arvioida annotaatioiden laatua. Vertailu tehdään genomi-, eksoni- ja nukleotiditasolla. Jokaisella tasolla lasketaan vertausta kuvaava arvo, esimerkiksi nukleotiditasolla jokaiselle referenssin kanssa linjattavalle geenille lasketaan herkkyys (sensitivity) ja tarkkuus (specificity) ja niiden arvoilla lasketaan f-score. Aineistona käytettiin neljää metsämansikan (*Fragaria vesca*) genomin eri versiota ja niiden kuutta annotaatiota. Ne ladattiin Genome Database for Rosaceae tietokannasta, joka on ruusukasveihin erikoistunut genomi-tietokanta.

Annotaatioista laskettujen laatuarvojen ja referenssiin vertausta kuvaavan arvon korrelaatio-kerroin oli useassa tapauksessa pieni, mutta luotettava, koska kaksisuuntainen p-arvo oli minimaalinen. Korrelaatiokertoimet olivat suurempia, kun tutkittiin proteiinien homologiaan perustuvia laatumittareita. Rakenteisiin perustuvien laatumittarien keskiarvon ja f-scoren välinen korrelaatiokerroin sai pienempiä arvoja, jos tutkittava annotaatio sai hyvän f-scoren arvon.

Tulokset tukevat näkemystä, että valitut rakenteisiin perustuvat laatumittarit eivät sovellu korkealaatuisten annotaatioiden laadunarviointiin. Niiden mahdollinen käyttötarkoitus voisi olla huonolaatuisten annotaatioiden automaattinen löytäminen. Laatumittarit, jotka perustuivat geenin proteiinituotteen ja proteiinitietokannan vertailuun, vaikuttivat lupaavilta jatkotutkimuksen kohteilta.

Avainsanat — Nyckelord — Keywords

Bioinformatiikka, geeninennustus, evaluointi, genomi

Säilytyspaikka — Förvaringsställe — Where deposited

Maataloustieteiden maisteriohjelma, maataloustieteiden osasto

Muita tietoja — Övriga uppgifter — Further information

Ohjaaja: Professori Liisa Holm, bio- ja ympäristötieteellinen tiedekunta, HY

ABSTRACT

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Osasto — Sektion — Department
Faculty of Agriculture and Forestry	Department of Agricultural Sciences

Tekijä — Författare — Author

Antti Tuominiemi

Työn nimi — Arbetets titel — Title

On the evaluation of gene predictions: application to strawberry genomes

Oppiaine —Läroämne — Subject

Biotechnology

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
M.Sc. Thesis	1 November 2020	37

Tiivistelmä — Referat — Abstract

The sequencing methods used to study the genome of organisms have become cheaper, resulting in a significant increase in the amount of genomic data available. Knowing the nucleic acid sequence of the DNA does not tell much about an organism. Not without first annotating the genome, which means searching for the locations of genes and defining their products. The programs used for annotation make mistakes and their results must be evaluated in various ways. The vast amount of genomic data encourages fast production of new annotations and this can increase human made errors. Some annotation programs use gene databases, so the number of wrongly annotated genes they contain may increase in the future if the quality control of annotations is not improved.

This study examines correlation between selected quality measures and the quality of annotations. The quality metrics used can be divided into two basic types, the first one is based on the basic structures of genes and the second one on comparing the protein product of a gene against a protein database. The study assumes that comparison to a reference is a reliable way to assess the quality of annotations. The comparison is made at genome, exon and nucleotide levels. A single value describing the comparison is calculated at each level. For each gene aligned with a reference gene, sensitivity and specificity are calculated and used to make f-score at the nucleotide level. Four different versions of the wild strawberry (*Fragaria vesca*) genome and their six annotations were used as data. They were downloaded from the Genome Database for Rosacaea, which is a genome database specializing in rose plants.

The correlation coefficients calculated from quality metrics and f-scores were in several cases small but reliable because the p-value was minimal. Correlation coefficients were higher when quality metrics based on protein homology were examined. The correlation coefficient calculated from the mean of the structure-based quality metrics and the f-score received lower values if the studied annotation had a high f-score value.

These results detailed in this paper support the view that the selected structure-based quality metrics are not suitable for evaluation of high-grade annotations. They might possibly be used in automated detection of poor-quality annotations. Quality metrics based on protein homology appeared to be promising subjects for further research.

Avainsanat — Nyckelord — Keywords

Bioinformatics, gene prediction, evaluation, genome

Säilytyspaikka — Förvaringsställe — Where deposited

Master's Programme in Agricultural Sciences, Department of Agricultural Sciences

Muita tietoja — Övriga uppgifter — Further information

Supervisor: Professor Liisa Holm, Faculty of biological and enviromental sciences, UH

Sisällys

1 Johdanto

Eliöt siirtävät perintöaineksensa DNA:n muodossa jälkeläisilleen ja sukupolvien saatossa eliöiden lisäksi myös DNA muuttuu. Näistä pienistä muutoksista kumpuaa elämän monimuotoisuus. Ennen viime vuosituhannen loppupuolta emme pystyneet tutkimaan perintöainekseen sisällytettyä tietoa, mutta sitten kehitimme DNA sekvensoinnin. Sen avulla pystymme selvittämään eristetyn DNA:n emäsjärjestyksen, mutta uusimmatkaan sekvensointimenetelmät eivät ole erehtymättömiä ja voivat tehdä virheitä. Alun perin sekvensointi oli todella kallista, mutta kehityksen myötä hinnat ovat tulleet alas. Halpenemisen takia aikaisempaa useampi tutkimusryhmä on voinut osallistua eliöiden genomien sekvensointiin. Uudet ja mahdollisesti kokemattomat tutkijat lisäävät sekvensoinnin virhemahdollisuuksia. DNA:n emäsjärjestyksen selvittäminen ei yksinään kerro tutkittavan eliön toiminnasta paljonkaan, vaan sen lisäksi täytyy pystyä selvittämään sekvenssin toiminnalliset alueet. Geenit ovat perimän alueita, jotka eliö siirrostaa mRNA:ksi (lähetti-RNA). Aitotumallisissa eliöissä geenit rakentuvat säätelyalueista, tietoa sisältävistä eksoneista ja silmukoinnissa poistettavista introneista. Säätelyalueet voivat sijaita myös kaukana säätelemästään geenistä. Proteiinia tuottavassa geenissä ensimmäinen ja viimeinen eksoni sisältävät myös UTR:än (EI-koodaavan alueen), joka nimensä mukaisesti ei vaikuta tuotetun proteiinin aminohapposekvenssiin. Silmukoinnissa 99% poistettavista introneista alkavat emäsparilla GU ja loppuvat emäspariin AG. Tätä kutsutaan kanoniseksi silmukoinniksi (canonical splicing). Translaatiossa proteiinit alkavat aina metioniini aminohapolla, mutta eliö voi jälkimuokkauksella poistaa sen. Translaatiossa ribosomi tulkkaa mRNA:n emäs-järjestystä kolme emästä kerrallaan yhdeksi aminohapoksi kasvavaan proteiiniin. Tätä kolmikko jaottelua kutsutaan geenin lukukehykseksi. Translaation lopussa ribosomi kohtaa mRNA:ssa lopetuskodonin ja irtoaa siitä vapauttaen tuotetun proteiinin samalla.

Suuri osa genomitutkimuksesta keskittyy proteiineja tuottavien geenien löytämiseen tutkittavasta perimästä. Ohjelmia, joilla etsitään geenejä, kutsutaan geenien ennustus-ohjelmiksi ja niiden löytämiä mahdollisia geenejä kutsutaan geeniennusteiksi. Toinen termi geenien löytämiselle, nimeämiselle ja funktion määrittämiselle on annotointi. Annotaatio sanaa voidaan myös käyttää substantiivina tällöin se viittaa annotoituihin geeneihin. Tavallisesti geenien ennustusohjelmat raportoivat geeneistä vain eksonit ja

intronit. Niitä etsitään pääsääntöisesti kahdella tavalla tai niiden yhdistelmällä. Ensimmäisissä eli niin kutsutuissa ab initio -menetelmissä geenejä löydetään aikaisemmin luodulla matemaattisella mallilla, joka on harjoitettu aikaisemmilla löydöksillä tai eliön sukulaisilta löydetyillä geeneillä. Käytetyt matemaattiset mallit yritetään rakentaa siten, että ne havaitsevat sekvenssin sisäisiä signaaleja. Toisessa eli vertailevissa menetelmissä tutkittavalle sekvenssille linjataan toisia sekvenssejä esimerkiksi aikaisemmin löydettyjä geenejä tai kokeellisesti eristettyjen tai tietokannoissa olevien RNA:iden tai proteiinien sekvenssejä. RNA ja proteiini -dataa käytettäessä etuna on, että ne ovat rakentuneet vain eksoneista. Nämä keinot eivät ole erehtymättömiä eli tuotetussa annotaatiossa on virheitä. Koska osa geenien ennustusohjelmista käyttää geenitietokantoja ennusteiden tekoon, virheelliset geeniennusteet voivat yleistyä kyseisissä tietokannoissa. Näin ollen geeniennusteiden laadunvalvonta on todella tärkeää.

Tässä tutkimuksessa arvioimme geeniennusteita laatumittareilla, jotka perustamme geenien perustoimintoihin. Ennustettu geeni saa täydet pisteet, jos sen proteiinituote alkaa metioniini aminohapolla, se loppuu lopetuskodoniin, se ei sisällä eksonin sisäistä ylimääräistä lopetuskodonia käytetyssä lukukehyksessä, se ei sisällä lukukehyksen muutosta eksonin sisällä ja sen intronit noudattavat kanonista silmukointia. Lisäksi käytämme kolmea laatumittaa, jotka kuvastavat ennustetun geenin proteiinituotteen vertausta homologisiin proteiineihin.

2 Geenien ennustusmenetelmistä ja niiden laadunarvioinnista

2.1 Geenien ennustamisesta

Geenien ennustusohjelmat jaetaan perinteisesti kolmeen päätyyppiin ab initio, vertaileviin ja yhdistelmämenetelmiin. Ab initio -menetelmissä geeniennustus luodaan statistisilla menetelmillä tutkittavasta sekvenssistä. Vertailevissa menetelmissä sekvenssi linjataan erinäisissä tietokannoissa olevia aikaisemmin tunnettuja sekvenssejä vasten tai eristettyjä proteiini tai RNA sekvenssejä vasten. Yhdistelmämenetelmät nimensä mukaisesti yhdistävät edeltäviä menetelmiä paikaten näiden heikkoja puolia. Ab initio -menetelmät tavallisesti tuottavat myös vääriä ennusteita. Vertailevien menetelmien heikkona kohtana on geenien löytäminen, jotka eroavat merkittävästi tietokannoissa

olevista. Sivutamme puhtaat vertailevat geenien ennustusmenetelmät, sillä tutkittavista geeniennusteista yksikään ei ole tuotettu siten.

2.1.1 Ab initio -menetelmistä

GeneMark-ES oli ensimmäinen itseoppiva aitotumallisten geeniennusteita tuottava ohjelma. Eli se voi muokata sisäisiä painotuksiaan parantaen tuloksia. Pohjimmiltaan se tuottaa geeniennusteita käyttäen algoritmia, joka perustuu Markovin piilomalliin, joka rakentuu tiloista ja näiden välisistä siirtymätodennäköisyyksistä (Lomsadzea ym. 2005). GeneMark-ES+ täsmentää geeniennusteita yhdistämällä teoreettisesta mallinnuksesta saatuja ennusteita kokeellisten tulosten kanssa, kuten esimerkiksi transkriptomi sekvensoinnista (Shulaev ym. 2011a).

AUGUSTUS nimisen geenien ennustusohjelman toiminta pohjautuu Markovin piilo-malliin, jossa on mallinnettuna intronit, eksonit, geenien väliset alueet ja muita geenin toiminnalle tärkeitä alueita (Stanke ja Waack 2003). Aitotumallisten geeneissä on mahdollista, että tapahtuu vaihtoehtoista silmukointia eli yksi geeni voi tuottaa enemmän kuin vain yhdenlaista proteiinia. Tällöin geenissä on eksoneja, jotka eivät aina päädy lopputuotteeseen. AUGUSTUS geeniennustusohjelman asetuksia voidaan muokata, siten että se tuottaa vaihtoehtoisia eksonikokonaisuuksia geeneille (Stanke ym. 2006). Geeniennusteita voidaan tuottaa, joko aikaisemmin määritellyillä ennustus asetuksilla tai ohjelmaa voidaan harjoittaa, jotta sen algoritmin painotuksia saadaan hienosäädettyä. Harjoittaminen tehdään antamalla ohjelmalle sekvensoitu tutkittava genomi ja tutkittu transkriptomi, proteomi tai luotettuja annotoituja geenejä (Hoff ja Stanke 2013).

SNAP on perinteinen Markovin piilomalleja hyödyntävä geenien ennustusohjelma, mutta se oli yksi ensimmäisistä, jonka käyttäjä pystyi harjoittamaan toimimaan uuden eliön genomilla (Korf 2004). Artikkelissaan Korf kertoo merkittävistä parannuksista geeniennusteisiin, jotka hän sai aikaan harjoittamalla SNAP:pia muiden ennus-tusohjelmien tuloksilla. Yhtäkään hänen käyttämistään geenien ennustusohjelmista ei ollut harjoitettu tutkittaville eliöille.

2.1.2 Yhdistelmämenetelmistä

MAKER ohjelman avulla yhdistetään monesta eri lähteestä peräisin olevia laskennalliseen mallinnukseen perustuvia geeniennusteita ja kokeellisesta datasta saatuja todisteita (Holt ja Yandell 2011). Alun perin MAKER:issä oli sisäänrakennettuna SNAP niminen geenien ennustusohjelma, joka harjoitettiin CEGMA menetelmällä genomista löydetyillä biologisesti todella tärkeillä geeneillä. Sitten SNAP:illa ennustettiin osasta genomia geenejä, joista valittiin geenit, joiden todenperäisyyttä tukivat proteiinilinjaukset ja cDNA data. Näitä luotettavia geenejä käytettiin uudelleen harjoittamaan SNAP, jonka jälkeen vasta luotiin lopulliset geeniennusteet (Cantarel ym. 2008). Myöhemmässä versiossa MAKER2 geeniennustus voitiin suorittaa, joko AUGUSTU:ksella, GeneMarkilla tai SNAP:illa. Samalla kertaa ohjelmaan lisättiin myös tuki uudelleen annotoida genomeja antamalla sille syötteenä vanhoja geeniennusteita genomin ja kokeellisen datan ohella (Holt ja Yandell 2011). Ohjelmasta on myös tehty versio MAKER-P, joka on optimoitu kasveille. Siihen on lisätty mekanismeja, joilla voidaan tunnistaa pseudogeenit ja koodaamattomat RNA:t, eli transkriptomin mukana olevat RNA pätkät, jotka eivät tuota proteiineja (Campbell ym. 2014).

2.2 Geeniennusteiden laadunarvioinnista

2.2.1 Vertaus luotettavaan referenssiin

Teoriassa geeniennusteiden vertaaminen virheettömään referenssiin on keino, joka kertoo ennusteiden oikeellisuuden absoluuttisesti, mutta käytännössä tällaista referenssiä ei ole. Näin ollen referenssin luotettavuus on tämän laadun arviointimenetelmän tulosten hyödyllisyyden kannalta keskeisin muuttuja. Vertailu tehdään tavallisesti nukleotidi-, eksoni- ja proteiinitasolla (Burset ja Guigo 1996). Nukleotiditasolla tutkittavan ennusteen nukleotideille annetaan totuusarvo, joka on joko oikea positiivinen (true positive), väärä positiivinen (false positive), oikea negatiivinen (true negative) tai väärä negatiivinen (false negative). Jos nukleotidi on ennusteessa ja referenssissä osana eksonia se saa arvon oikea positiivinen. Jos nukleotidi on ennusteessa merkattu koodaavaksi, mutta referenssissä se ei ole osa eksonia, se saa arvon väärä positiivinen. Nukleotidin ollessa referenssissä ja ennusteessa merkattu eksonin ulkopuolelle se saa arvon oikea

negatiivinen. Referenssissä koodaavana ja ennusteessa koodaamattomaksi merkattu nukleotidi saa arvon väärä negatiivinen (taulukko 1).

Taulukko 1. Totuusarvotaulukko.

		Referenssi	
		Koodaava	Ei koodaava
Ennuste	Koodaava	Oikea Positiivinen	Väärä Positiivinen
	Ei koodaava	Väärä Negatiivinen	Oikea Negatiivinen

Näiden tietojen avulla voidaan laskea kaksi vertailua kuvaavaa arvoa, herkkyys (sensitivity) ja tarkkuus (specificity). Herkkyys on koodaavien nukleotidien osuus, jotka on onnistuneesti ennustettu koodaaviksi. Tarkkuus on koodaaviksi ennustettujen nukleotidien osuus, jotka onnistuneesti ennustettu koodaaviksi. Eksonitasolla eksoni on oikea positiivinen vain ja vain jos sen alku- ja loppukohdat ovat identtiset referenssissä olevan sitä vastaavan eksonin kanssa. Täten eksonitasolla on myös ennustettuja eksoneja, jotka sijaitsevat suurin piirtein samassa kohdassa kuin referenssissä, mutta niitä ei lasketa oikeiksi positiivisiksi. Näin ollen eksoni voi olla puuttuva eli ennusteessa ei ole referenssissä olevan eksonin kohdalla mitään. Näillä voidaan laskea puuttuvien eksonien osuus referenssin eksoneista. Eksoni voi myös olla väärä eli ennustetulla eksonilla ei ole minkäänlaista vastaavuutta referenssissä, joten voidaan laskea väärien eksonien osuus kaikista ennustetuista eksoneista. Proteiinitasolla verrataan geeniennusteen teoreettisen proteiinituotteen aminohapposekvenssiä sitä vastaavan referenssigeenin aminohappo-sekvenssiin. Ne linjataan keskenään ja lasketaan oikein linjattujen aminohappojen osuus koko linjauksen pituudesta eli aminohappoidenttisyys.

2.2.2 Tietokantoihin vertailevista laadunarviointikeinoista

Eräs tapa tutkia geeniennusteiden luotettavuutta on keskittyä tiettyihin geeneihin, jotka oletetaan olevan edustettuina tutkittavassa organismissa. CEGMA (Core Eukaryotic Genes Mapping Approach) -menetelmän tarkoitus on löytää syötetystä aitotumallisen genomista geenejä, jotka tuottavat biologisesti elintärkeitä proteiineja (Parra ym. 2007).

Menetelmän alkulähtökohdat ovat tutkittava genomi ja kokoelma proteiineja, jotka on valittu niiden yleisyyden takia. Oletuksena on, että osa kokoelman proteiineista tarvitaan johonkin niin tärkeään prosessiin eliössä, että niissä ei tapahdu eliöiden tai edes eliöryhmien välillä suurta muuntelua. Proteiineille etsitään genomista vastaavuuksia ja tuotetaan geeniennusteet, joiden teoreettiset proteiinituotteet eivät eroa liian paljon alun perin sitä vastaavasta proteiinista. Näitä geeniennusteita käytetään vielä Markovin piilomallia käyttävän geenien ennustusohjelman harjoittamiseen, joka sen jälkeen löytää lopulliset geenit (Parra ym. 2007). Myöhäisemmässä julkaisussa (Parra ym. 2009) menetelmän kehittäjät rajaavat genomista etsittävää proteiinijoukkoa sellaisiin, joita tuottavat geenit esiintyvät yleensä yksittäisinä kappaleina genomissa. He myös tuovat ilmi, että proteiinit, joita vastaavia geenejä ei löydetty tutkittavasta genomista, kertovat puutteista genomin sekvensoinnissa. Vertaamalla CEGMA menetelmällä löydettyjä geenejä kattavampien geenien ennustusmenetelmien tuloksiin voidaan niiden laatua arvioida. CEGMA:n kehitys ja tuki lopetettiin vuonna 2015 (Korf 2015).

BUSCO, eli benchmarking sets of universal single-copy orthologs, ovat kokoelma geenejä, jotka on valikoitu kokoelmaan, jos niiden vastine on ollut edustettuna ainakin 90%:lla tutkittavan eliöryhmän lajeista yksittäisenä kappaleena perimässä (simao ym. 2015). Niiden löytämiseen on käytetty ortologien tietokantaa nimeltä orthodb (Kriventseva ym. 2018). Kaksi geeniä ovat ortologeja, jos ne ovat toisilleen sukua olevissa eliöissä ja ovat kehittyneet yhteisen kantaisän samasta geenistä. BUSCO:n neljännessä versiossa on geenikokoelmia bakteereille, arkeille, alkueläimille, sienille, kasveille ja eläimille, sekä osalle niiden alaryhmistä. BUSCO analyysi ei ota kantaa yksittäisten ennustettujen geenien validiteettiin, muuten kuin siinä ääritapauksessa, että kyseinen geeni on BUSCO setissä olevan geenin ortologi, vaan arvioi annotaation luotettavuutta kokonaisuutena.

Ennustettujen geenien teoreettisten proteiinituotteiden analysointi on yksi keino saada tietoa geeniennusteiden luotettavuudesta. GeneValidator työkalun avulla voidaan verrata niitä useissa tietokannoissa oleviin proteiineihin (Dragan ym. 2016). Se etsii BLAST:in avulla tietokannoista tutkittavaa teoreettista tuotetta eniten muistuttavat proteiinit ja pisteyttää jokaisen ennusteen neljällä eri tavalla, jotka ovat pituus, kattavuus, säilyneet alueet ja menetelmä, joka tutkii, onko tarkasteltava proteiini yhdistelmä kahdesta eri tietokannan proteiinista.

3 Tutkimuksen tavoitteet

Työn tarkoituksena on selvittää, voidaanko geenien perusrakenteisiin ja proteiinien homologiaan perustuvilla laatumittareilla arvioida geeniennusteiden laatua, joko yksittäisen geenin osalta tai annotaation mittakaavassa. Alkulähtökohtana on, että geeniennusteiden vertaaminen luotettavaan referenssiin kertoo niiden luotettavuudesta. Tutkimushypoteesina esitetään, että referenssiin vertauksesta kertova arvo korreloi valittujen laatumittarien keskiarvon kanssa.

4 Aineisto ja menetelmät

4.1 Annotaatiot ja assemblyt

Tutkittavat geeniennusteet ja niiden tuottamiseen käytetyt sekvensoidut genomit ladattiin GDR (Genome Database for Rosaceae) tietokannasta, joka on ruusukasvien heimoon erikoistunut genomitietokanta (Jung ym. 2014). Metsämansikan (*Fragaria vesca*) sekvensoituja genomeja oli neljä, joista oli tuotettu seitsemän annotaatiota (Taulukko 2). Datasetit v1.1.a2 ja v2.a2 oli tuotettu MAKER2 ohjelmalla hyödyntäen transkriptomidataa, edellisiä annotaatioita, sekä SNAP, Augustus ja GeneMark ennustusohjelmia. Uusin v4.a2 oli tuotettu yhdistämällä EvidenceModeler ohjelmalla Augustuksen ja MAKER2 ohjelmien tuottamat geeniennusteet. Datasetit v1.1.a1 ja v2.a1 oli tuotettu siirrostamalla edellinen annotaatio uudemmalle sekvensoidulle genomille. Annotaatioista vanhin oli vuodelta 2010 ja uusin vuodelta 2019. Vähiten ennustettuja geenejä oli ensimmäisessä annotaatiossa v1.0 ja eniten uusimmassa v4.a2. Neljä annotaatiota sisälsivät geenejä, joille oli ennustettu useita mRNA:ita, v1.1.a2, v2.0.a2, v4.0.a1 ja v4.0.a2. Uusin annotaatio v4.a2 valittiin referenssiksi, koska sen julkaisuartikkelin (Li ym. 2019) mukaan annotaation BUSCO complete score on 98,1%, eli se sisältää kokonaisina 98,1% tutkituista BUSCO geeneistä. Kaikki annotaatiot olivat GFF3 tiedostomuodossa, eli ne eivät sisällä sekvenssejä vaan kertovat geenin, mRNA:n tai eksonin aloituskohdan ja lopetuskohdan erillisessä genomitiedostossa.

Taulukko 2. Tutkittavat geeniennusteet, niiden sisältämien yksittäisten ennustettujen geenien määrä, datasetin tuotantotapa, julkaisuartikkelin ilmoittama BUSCO complete score, tähdellä merkityt BUSCO complete scoret ovat itse ajetulla BUSCO:lla tuotettuja ja lähdeviittaus.

Annotaatio	Geenien määrä	Tuotantotapa	BUSCO c	lähde
v4.0.a2	34009	EvidenceModeler, Augustus, MAKER2	98,1%	(Li ym. 2019)
v4.0.a1	28588	MAKER-P	95,0%	(Edger ym. 2018)
v2.0.a2	33538	MAKER2	95,7%	(Li ym. 2018)
v2.0.a1	33673	siirrostus v1.1.a1	88,9%	(Tennessen ym. 2014)
v1.1.a2	33507	MAKER2	87,8%*	(Darwish ym. 2015)
V1.1.a1	32531	siirrostus v1.0	53,3%*	(Shulaev ym. 2011b)
v1.0	28540	GeneMark-ES+	43,2%*	(Shulaev ym. 2011a)

4.2 Annotaatioiden siirrostus

Kaikki ohjelmat, jotka on tuotettu työtä varten, on ohjelmoitu Perl ohjelmointikielellä, jos asiasta ei erikseen mainita. Koska annotaatiot oli tuotettu neljästä eri genomin versiosta, täytyi annotaatiot ensin siirrostaa samalle versiolle. Vanhemmat versiot siirrostettiin uusimpaan v4 sekvensoituun genomiin. Tähän käytettiin CrossMap ohjelmaa, joka tarvitsi syötteikseen siirrostettavan GFF tiedoston ja chain tiedoston. Chain tiedosto kertoo ohjelmalle mitkä kohdat lähde sekvenssissä linjautuvat kohde sekvenssin kanssa. Chain tiedosto tuotettiin noudattaen Ari Löytynojan kurssi-materiaalissaan kuvaamaa menetelmää (Löytynoja 2018). CrossMap tuotti tuloksena kaksi tiedostoa, joista toisessa olivat siirrostetut geeniennusteet ja toisessa geeniennusteet, joita se ei voinut siirrostaa. CrossMap ajettiin myös toiseen suuntaan, eli referenssi siirrostettiin vanhemmille sekvensoiduille genomeille ja tulostiedosto, joka kertoo mitkä referenssin geenit eivät siirrostuneet, otettiin talteen.

Huomattiin, että CrossMap oli siirrostanut useita satoja geeniennusteita kahteen eri paikkaan (taulukko 3), täten luoden monistuneita geenejä, joilla oli sama nimi. Luotiin

ohjelma, joka etsii geenit, joilla on samat nimet, ja poistaa geenit, joilla oli vähemmän kappalemäärällisesti eksoneja. Pienelle osalle siirrostetuista geeniennusteista CrossMap oli tehnyt toisenlaisen virheen ja siirrostanut ne, vaikka tämä muutti suuresti niiden sisällä olevien intronien pituutta, eli nämä siirrokset ovat hyvin luultavasti virheellisiä. Tämän takia tuotettiin kaksi pientä ohjelmaa. Ensimmäisen, joka vertaa siirrostettujen geeniennusteiden pituutta siirrostamattomiin geeniennusteisiin, ja toisen, joka käyttää edellä mainitun ohjelman tuloksia ja poistaa siirrostettujen geeniennusteiden joukosta ne, joiden pituus oli alle kolmanneksen tai yli kolme kertaa niiden pituus ennen siirrostusta. Ohjelmat käytettiin kaikkiin vanhemmilla sekvensoiduilla genomeilla tuotettuihin geeniennusteisiin.

Taulukko 3. Tutkittavista dataseteistä eri vaiheissa poistettujen geeniennusteiden määrät.

	alkuperäinen määrä	crossmappauksessa poistuneet	crossmappauksessa monistuneet	väärän kokoiset	poistettavia koska eksonit päällekkäisiä	poistetun referenssin kanssa linjautuneet	poistettu koska exonerate ei linjannut	yhteensä poistettavien määrä	jäljellä
v4.a1	28588	0	0	0	0	0	6	6	28582
v2.a2	33538	14	222	210	7	2	4	459	33079
v2.a1	33673	27	316	246	5	2	4	600	33073
v1.1.a2	33507	2146	719	209	14	4	22	3114	30393
v1.1.a1	32531	1341	244	207	6	2	5	1805	30726
v1.0	28540	7	216	188	6	5	4	426	28114

4.3 Referenssiin vertaus

Luotiin ohjelma, jolla poistettiin referenssistä vaihtoehtoiset mRNA:t, jotta myöhemmin referenssiin vertauksessa tutkittavilla geeneillä olisi vain yksi vaihtoehto referenssissä. Ohjelma valitsi jatkoon mRNA:n, jonka eksonien yhteenlaskettu pituus oli suurin. Jos kaksi tai enemmän vaihtoehtoja olivat yhtä pitkiä, ohjelma valitsi ensimmäisen. Myöhäisemmässä vaiheessa huomattiin, että hyvin harvat ennustetut geenit sisältävät saman mRNA:n alaisia eksoneja, jotka ovat osittain päällekkäin, joka on selvästi biologisesti mahdotonta. Suunniteltiin ja suoritettiin ohjelma, joka etsi kyseiset mahdottomat geeniennusteet ja poisti ne (taulukko 3).

Seuraavaksi kasattiin ohjelma, joka vertaa siirrostettujen geeniennusteiden alku- ja loppukohtia valitun referenssin geenien alku- ja loppukohtiin. Ohjelma raportoi, mitkä

geeniennusteen yksittäiset geenit jakavat yhteisiä kohtia referenssin geenien kanssa eli limittäiset geenit, sekä mitkä ennustetut geenit eivät linjaudu referenssin minkään geenin kanssa eli väärät geenit. Se ilmoittaa myös mille referenssin geeneille ei ole ennustettu vastinetta eli puuttuvat geenit. Ohjelma ajettiin kaikille geeniennusteille. Näistä tuloksista poistettiin limittäiset geenit ja puuttuvat geenit, joiden referenssigeeniä ei voitu siirrostaa kyseisen geeniennusteen tuottamiseen käytettyyn sekvensoituun genomiin (taulukko 3).

Tehtiin kaksi ohjelmaa, joista ensimmäinen vertaa edellisen ohjelman tuloksen limittäisiä geenejä niiden referenssi vastinkappaleisiin ja ilmoittaa niiden sisäiset limittäiset, väärät ja puuttuvat eksonit ja toinen laskee kyseisille limittäisille geeneille herkkyyden (1), tarkkuuden (2), puuttuvien eksonien osuuden referenssin eksoneista (3) ja väärien eksonien osuuden ennustetuista eksoneista (4). Ohjelmilla laskettiin geeniennusteille nämä suureet.

$$Sn = \frac{o}{r}, \quad (1)$$

$$Sp = \frac{o}{q}, \quad (2)$$

$$Sn' = \frac{p}{r}, \quad (3)$$

$$Sp' = \frac{v}{q}, \quad (4)$$

joissa
Sn = herkkyys
Sp = tarkkuus
Sn' = herkkyys puuttuville eksoneille
Sp' = tarkkuus väärille eksoneille
o = oikein ennustettu eksoni
r = referenssigeenin eksonien määrä
q = geeniennusteen eksonien määrä
p = puuttuvien eksonien määrä
v = väärien eksonien määrä.

Koska kolmessa tutkittavassa annotaatiossa esiintyy vaihtoehtoisia mRNA:ita tehtiin ohjelma, joka valitsee edellisen ohjelman tulosten mukaan jokaiselle referenssin kanssa

limittäiselle ennustetulle geenille parhaan mRNA:n siten, että ohjelma laskee mRNA:lle f-scoren, joka on tarkkuuden ja herkkyyden harmoninen keskiarvo (5). Tällä valittiin limittäisille geeneille luotettavin mRNA.

$$f = \frac{2 * Sn * Sp}{Sn + Sp}, \quad (5)$$

jossa
f = f-score
Sn = herkkyys
Sp = tarkkuus.

Suunniteltiin ohjelma, joka laskee limittäisille geeneille nukleotiditasolla herkkyyden ja tarkkuuden. Ensimmäisellä ajolla hyvin harvat geenit tuottivat tuloksia, jotka olivat mahdottomia siten, että niiden tarkkuus tai herkkyys oli yli 1. Ongelmaksi paljastui muutama ennustettu geeni, jonka mRNA:n alaiset eksonit olivat päällekkäisiä. Kyseiset biologisesti mahdottomat ennustetut geenit poistettiin tutkittavien geenien ryhmästä. Tämän jälkeen ohjelman tulokset eivät enää olleet mahdottomia.

Viimeisenä ohjelmista, jotka ovat vastuussa referenssiin vertaamisesta, tehtiin ja ajettiin genomitason herkkyys ja tarkkuus laskin, joka laskee niiden lisäksi myös herkkyyden puuttuville geeneille ja tarkkuuden väärille geeneille. Se hyväksyy geenin oikeaksi vain, jos referenssi on täysin identtinen ennusteen kanssa eli, jos sen alku ja loppu ovat identtiset referenssissä ja ennusteessa, sekä sen eksonitason herkkyys ja tarkkuus ovat molemmat yksi.

4.4 Laatumittarien laskeminen

Seuraavana vaiheena oli laatumittarien laskemiseen käytetyn ohjelman rakentaminen. Valitut geeniennusteiden sisäiset laatumittarit laskettiin eksonien sisäisien lukukehyksien siirtymien määrästä (6), kanonista silmukointia noudattavien intronien määrästä (7), eksonien sisäisten käytetyssä lukukehyksessä olevien stop kodonien määrästä (8), metioniini aminohapolla alkamisesta ja stop kodoniin loppumisesta.

$$Ls = \frac{1}{1+ Fs}, \quad (6)$$

$$Ks = \frac{Ca}{Im}, \quad (7)$$

$$Se = \frac{1}{1+Ls}, \quad (8)$$

joissa

Ls = lukukehyksen siirtymiä kuvaava laatumittari

Ks = kanonista silmukointia kuvaava laatumittari

Se = lukukehyksessä olevia eksonin sisäisiä stop kodoneja kuvaava laatumittari

Fs = eksonien sisäisten lukukehyksen siirtymien määrä

Ca = kanonista silmukointia noudattavien intronien määrä

Im = intronien määrä

Ls = Lukukehyksessä olevien eksonien sisäisten stop kodonien määrä.

Teoreettisen proteiinin vertailusta sen homologi proteiineihin monisekvenssilinjauksella voitiin laskea kolme vertailuarvoa, jotka kuvaavat sen osuvuutta. Ensimmäinen kertoo identtisesti linjautuvien aminohappojen osuuden tutkittavasta proteiinista (pide) (9). Monisekvenssilinjauksessa alussa ja lopussa voi olla ulkonemia linjauksessa. Alun ulkonemilla lasketaan toinen (nter) (10) ja lopun ulkonemilla kolmas (cter) (11) vertailuarvo. Toinen ja kolmas vertailuarvo ovat pieniä, jos teoreettinen proteiini vertautuu pidempiin tai lyhyempiin todellisiin proteiineihin.

$$P = lpide * qcov, \quad (9)$$

$$n = \frac{1}{1 + \frac{Sfrom-1+Qfrom-1}{10}}, \quad (10)$$

$$c = \frac{1}{1+ \frac{Sseqlen-Sto+Qseqlen-Qto}{10}}, \quad (11)$$

joissa

P = pide

n = nter

c = cter

lpide = SANS ohjelman raportoima pide, joka on laskettu linjauksen pituuden mukaan

qcov = linjauksen osuus tutkittavan proteiinin pituudesta
Sfrom = linjauksen alkukohta homologisessa proteiinissa
Qfrom = linjauksen alkukohta tutkittavassa proteiinissa
Sseqlen = homologisen proteiinin pituus
Sto = linjauksen loppukohta homologisessa proteiinissa
Qseqlen = tutkittavan proteiinin pituus
Qto = linjauksen loppukohta tutkittavassa proteiinissa.

Ensin geeniennusteillä täytyi tuottaa ennustetun teoreettisen proteiinin aminohapposekvenssi. Käytettiin tutkimusryhmän sisäisessä käytössä olevaa Perl:illä tehtyä pikkuohjelmaa, joka tuotti proteiinisekvenssit annetusta genomista ja geeniennustus tiedostosta. Ohjelma ajettiin alkuperäisillä GDR tietokannasta ladatuilla tiedostoilla, jotta minimoitiin tietojenkäsittelyssä mahdollisesti tapahtuneita virheitä.

Metioniini aminohapolla alkamisen ja stop kodoniin loppumisen ohjelma pystyi tarkistamaan suoraan teoreettisen proteiinin aminohapposekvenssistä. Muut olivat helposti saatavilla linjaamalla tuotettu teoreettinen proteiini Exonerate sekvenssi-linjaimella sen tuottanutta genomin kohtaa vasten. Exoneraten tuloksesta saatiin helposti muut sisäiset laatumittarit tutkimusryhmältä saadulla Perl-moduulilla. Ohjelma linjaa teoreettiset proteiinit SANS ohjelmalla homologi tietokantaa vastaan ja laskee tuloksista monisekvenssilinjauksen kymmenelle parhaalle edellä mainitut kolme vertailuarvoa ja laskee niiden keskiarvot.

Ohjelma ajettiin geeniennusteille ja huomattiin, että todella harvassa tapauksessa Exonerate ei pystynyt linjaamaan teoreettista proteiinia sen tuottaneeseen sekvenssiin. Kyseiset geenit poistettiin tutkittavien geenien joukosta (taulukko 3).

Viimeisenä kasattiin ohjelma, joka koosti geeniennusteiden yksittäisten geenien referenssiin vertauksesta kertovat arvot ja lasketut laatumittarit saman tulostiedoston alle. Se myös laski keskiarvon kaikista käytetyistä laatumittareista ja erillisen keskiarvon sisäisistä laatumittareista sekä keskiarvon monisekvenssilinjauksesta kertovista vertailuarvoista.

5 Tulokset ja niiden tarkastelu

5.1 Genomitaso

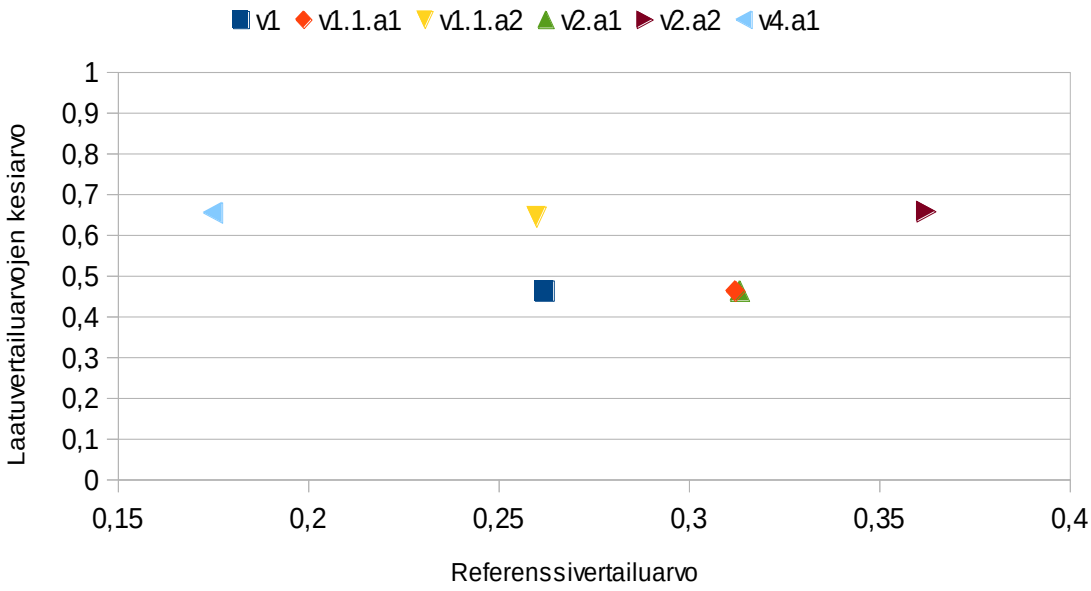
Genomitasolla kaikkien datasettien referenssivertailuarvo (12), joka on kaikkien neljän vertausta kuvaavan suureen harmoninen keskiarvo, oli välillä 0,17 – 0,36 (kuva 1). Käytetty laatuvertailuarvo on laatumittarien keskiarvo. Laatuvertailuarvot olivat välillä 0,46 – 0,66. Suurimman referenssivertailuarvon sai v2.a2. Se sai myös suurimman laatuvertailuarvon. Pienin referenssivertailuarvo oli uusimmalla ei referenssinä käytetyllä geeniennusteella v4.a1. Pienin laatumittarien keskiarvo oli geeniennusteella v1.1.a2. Kaikkien tutkittujen geeniennusteiden genomitason herkkyys ja tarkkuus olivat alle 0.25. Suurimman referenssivertailuarvon saanut geeniennuste sai myös suurimmat herkkyys ja tarkkuus arvot.

$$Rv = \frac{4}{\frac{1}{Sn}+\frac{1}{Sp}+\frac{1}{1-Sn'}+\frac{1}{1-Sp'}}, \qquad (12)$$

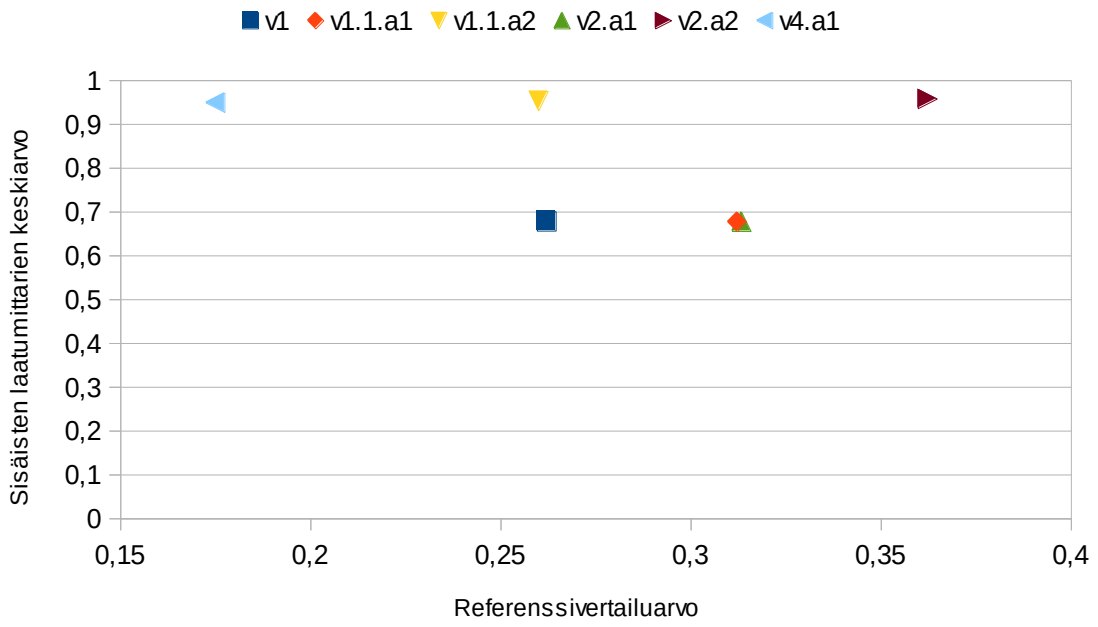
jossa
Rv = referenssivertailuarvo
Sn = herkkyys
Sp = tarkkuus
Sn’ = herkkyys puuttuville geeneille
Sp’ = tarkkuus väärille geeneille.

Genomitasolla tulokset jakautuivat laatuvertailuarvonsa mukaan selkeästi kahteen ryhmään. Datasetit v1.1a2, v2.a2 ja v4.a1 muodostivat ryhmän, jotka olivat saaneet toisiaan lähellä olleet korkeahkot laatuvertailuarvot. Loput kolme annotaatiota v1.0, v1.1.a1 ja v2.a1 muodostivat oman pienemmän laatuvertailuarvon ryhmänsä (kuva 1). Mielenkiintoisesti MAKER-ohjelmaperheellä tuotetut datasetit muodostivat edellä mainitun kolmen suuremman laatuvertailuarvon ryhmän. Referenssivertailuarvoilla ei ollut havaittavissa selkeää ryhmittymistä. Aikaisemmassa kappaleessa mainittu havainto, että korkeimman referenssivertailuarvon saanut datasetti sai myös suurimman laatuvertailuarvon, ei vaikuta merkittävältä, kun ottaa huomioon, että puolet dataseteistä saivat käytännössä yhtä suuren laatuvertailuarvon.

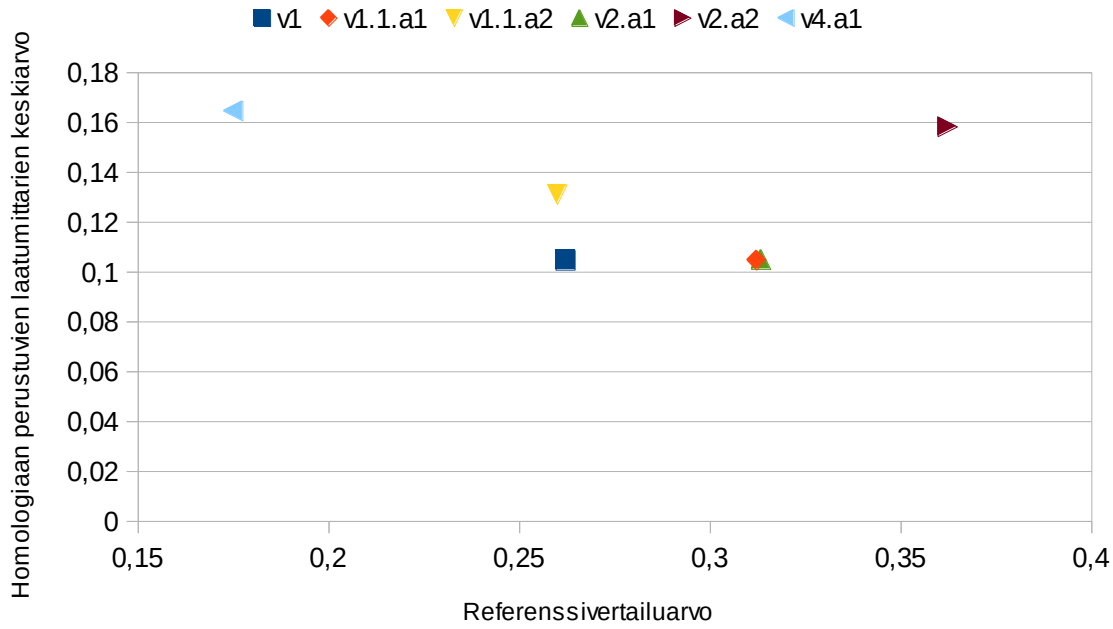
Laatuvertailuarvo on keskiarvo kahdeksasta laatumittarista, joista viisi on geeniennusteesta laskettavia (vastaisuudessa viittaamme näihin laatumittareihin sisäisinä laatumittareina) ja kolme proteiinien homologiaan perustuvia. Jos tarkastelemme geeniennusteiden sisäisten laatumittarien keskiarvoa verrattuna referenssivertailuarvoon (kuva 2) huomaamme, että datasetit jakautuvat kahteen samaan ryhmään kuin aikaisemmin. Jos taasen vertaamme kolmen proteiinien homologiaan perustuvan laatumittarin keskiarvoa referenssivertailuarvoon (kuva 3) sama jako on havaittavissa. Nämä tulokset eivät tue tutkimushypoteesiä.



Kuva 1. Datasettien Laatuvertailuarvojen keskiarvot verrattuna Referenssivertailuarvojen keskiarvoihin.



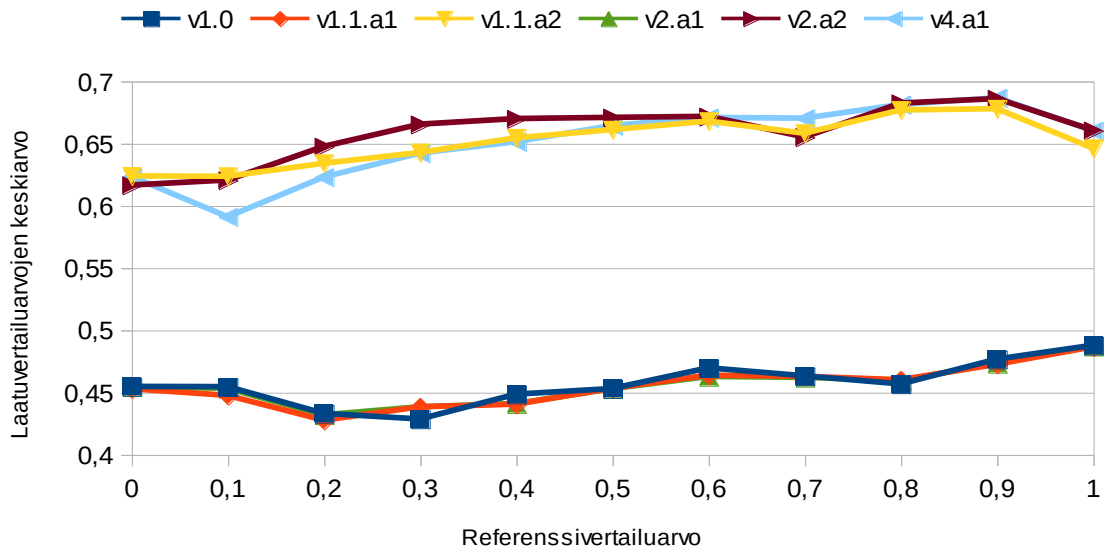
Kuva 2. Datasettien sisäisten laatuvertailuarvojen keskiarvot verrattuna referenssivertailuarvojen keskiarvoihin.



Kuva 3. Datasettien proteiinien homologiaan perustuvien laatuvertailuarvojen keskiarvo verrattuna referenssivertailuarvojen keskiarvoihin.

5.2 Eksonitaso

Jos vertaamme eksonitasolla referenssivertailuarvoa ja laatuvertailuarvoa keskenään jakautuvat datasetit kahteen ryhmään, joissa molemmissa on 3 datasettiä (kuva 2). Suuremman laatuvertailuarvojen keskiarvon ryhmään kuuluivat datasetit v1.1.a2, v2.a2 ja v4.a1. Vertailusta on havaittavissa pientä positiivista korrelaatiota referenssi-vertailuarvon ja laatuvertailuarvon välillä. Annotaatiolla v4.a1 oli suurin korrelaatiokerroin referenssivertailuarvon ja laatuvertailuarvon välillä (taulukko 4).

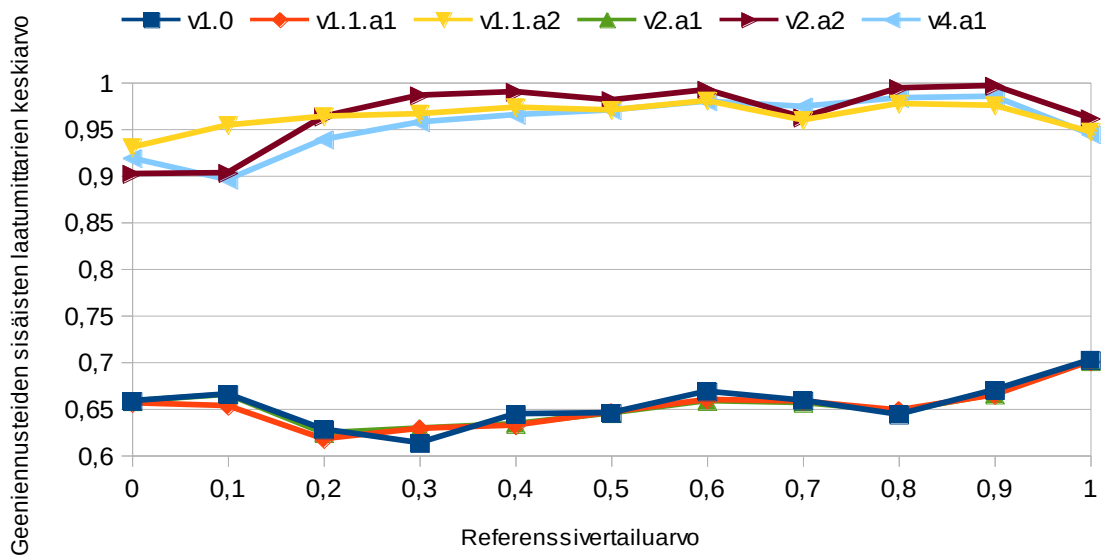


kuva 4. Eksonitason referenssivertailuarvo suhteessa kyseisen referenssivertailuarvon saaneiden geeniennusteiden laatuvertailuarvojen keskiarvoon.

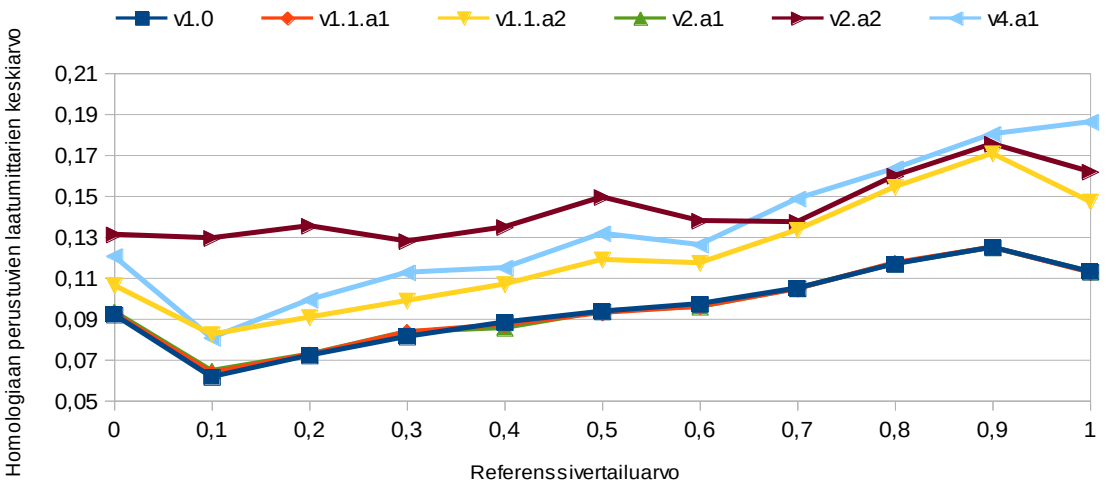
Taulukko 4. Referenssivertailuarvon, laatuvertailuarvon, geenien sisäisten laatumittarien ja proteiinien homologiaan perustuvien laatumittarien väliset korrelaatiokertoimet ja kaksisuuntaiset p-arvot.

		laatuvertailuarvo		sisäiset		homologiaan perustuvat	
annotaatio	otoskoko	korrelaatio	p-arvo	korrelaatio	p-arvo	korrelaatio	p-arvo
v4.a1	27733	0,2370	0,0000	0,0936	0,0000	0,1164	0,0000
v2.a2	30937	0,2240	0,0000	0,0938	0,0000	0,0445	0,0000
v2.a1	31957	0,0687	0,0000	0,0552	0,0000	0,0763	0,0000
v1.1.a2	30845	0,1842	0,0000	0,0624	0,0000	0,0693	0,0000
v1.1.a1	29587	0,0709	0,0000	0,0574	0,0000	0,0791	0,0000
v1.0	24132	0,0706	0,0000	0,0566	0,0000	0,0679	0,0000

Laatuvertailuarvon ja referenssivertailuarvon välinen heikko korrelaatio on havaittavissa, jos vertaamme referenssivertailuarvoa geeniennusteiden sisäisten laatumittarien keskiarvoon (kuva 5), sekä homologiaan perustuvien laatumittarien keskiarvoon (kuva 6). Sisäisten laatumittarien keskiarvossa voimme todeta selkeästi annotaatioiden



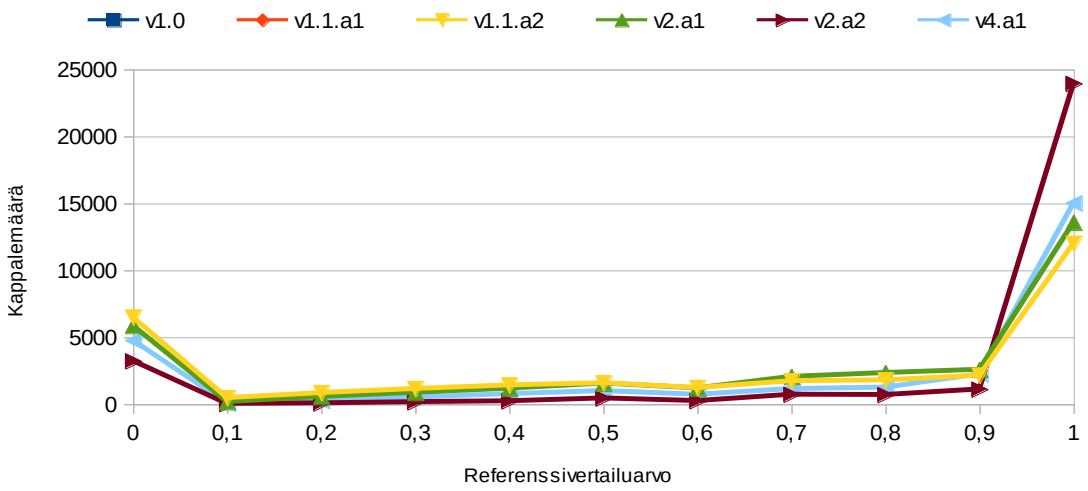
Kuva 5. Eksonitason referenssivertailuarvo suhteessa kyseisen referenssivertailuarvon saaneiden ennustettujen geenien geeniennusteiden sisäisten laatumittarien keskiarvoon.



Kuva 6. Eksonitason referenssivertailuarvo suhteessa kyseisen referenssivertailuarvon saaneiden geeniennusteiden homologiaan perustuvien laatumittarien keskiarvoon.

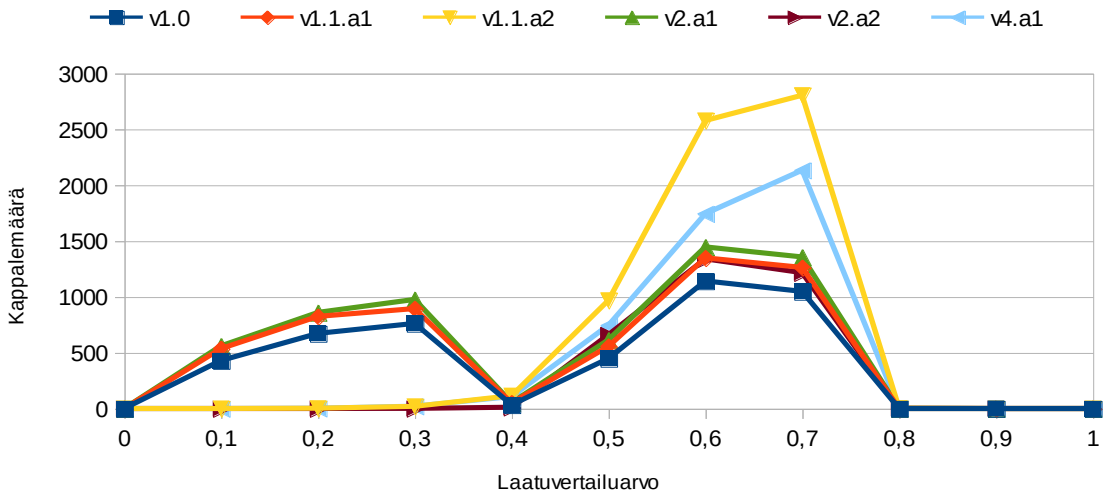
jakautuvan aikeisemmin mainittuihin pienemmän ja suuremman laatumittarien keskiarvon ryhmiin. Homologiaan perustuvien laatumittarien keskiarvojen vaihteluväli on todella pieni, mutta silti tutkimushypoteesiä tukeva korrelaatio on selkeästi havaittavissa.

Eksonitasolla jos ryhmittelemme geeniennusteet niiden referenssivertailuarvon (12) mukaan huomaamme, että kaikissa dataseteissä oli kappalemääräisesti eniten geeniennusteita, jotka saivat referenssivertailuarvon hyvin lähellä nollaa tai yhtä (kuva 7). Näin ollen tarkastelemme lähemmin kyseisiä geeniennusteita. Tutkittaessa geeniennusteita, joiden referenssivertailuarvo oli lähellä nollaa, osalla annotaatioista ei ollut geeniennusteita, joiden laatuvertailuarvo olisi ollut pieni (kuva 8). Tarkasteltaessa geeniennusteita, joiden referenssivertailuarvo oli lähellä yhtä, kappalemäärällisesti eniten ison laatuvertailuarvon omaavia geeniennusteita oli datasetissä v2.a2 (kuva 9).

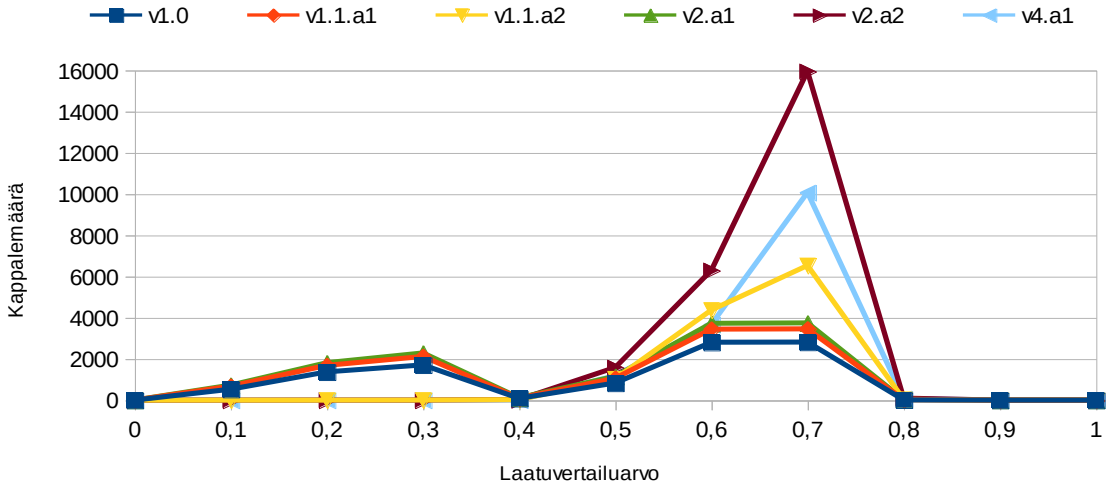


Kuva 7. Eksonitason referenssivertailuarvon jakauma tutkittavissa annotaatioissa.

Kun tarkastelemme laatuvertailuarvojen jakaumaa geeniennusteissa, jotka olivat saaneet referenssivertailuarvon lähellä nollaa tai yhtä, huomaamme miksi datasetit jakautuvat selkeästi kahteen eri ryhmään (kuvat 8 ja 9). Dataseteissä v1.0, v1.1.a1 ja v2.a1 geeniennusteet ovat jakautuneet kahteen ryhmään, jotka voidaan erotella niiden laatuvertailuarvosta.



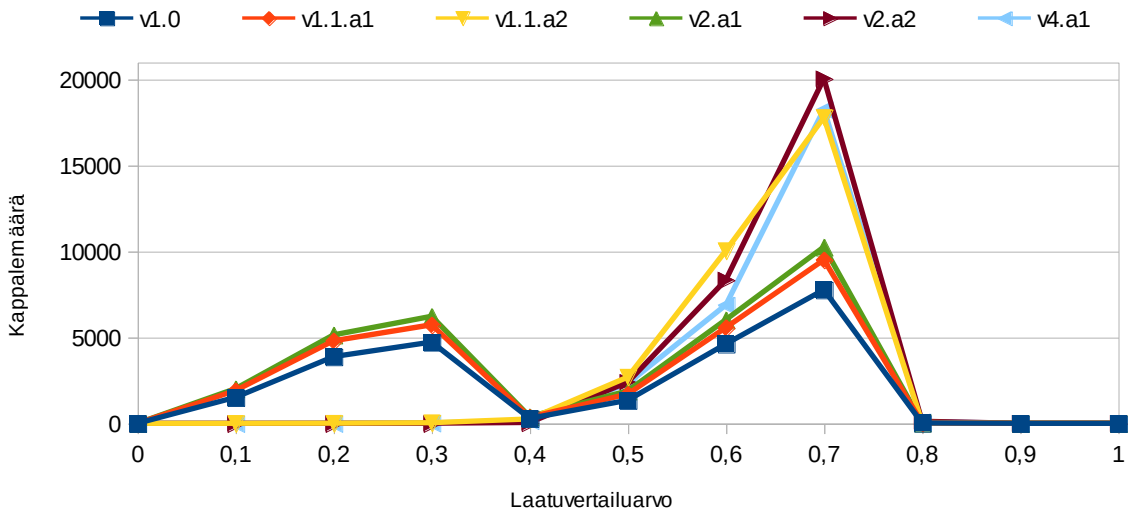
Kuva 8. Eksonitason laatuvertailuarvon jakauma tutkittavien annotaatioiden geeniennusteissa, joiden referenssivertailuarvo oli lähellä nollaa.



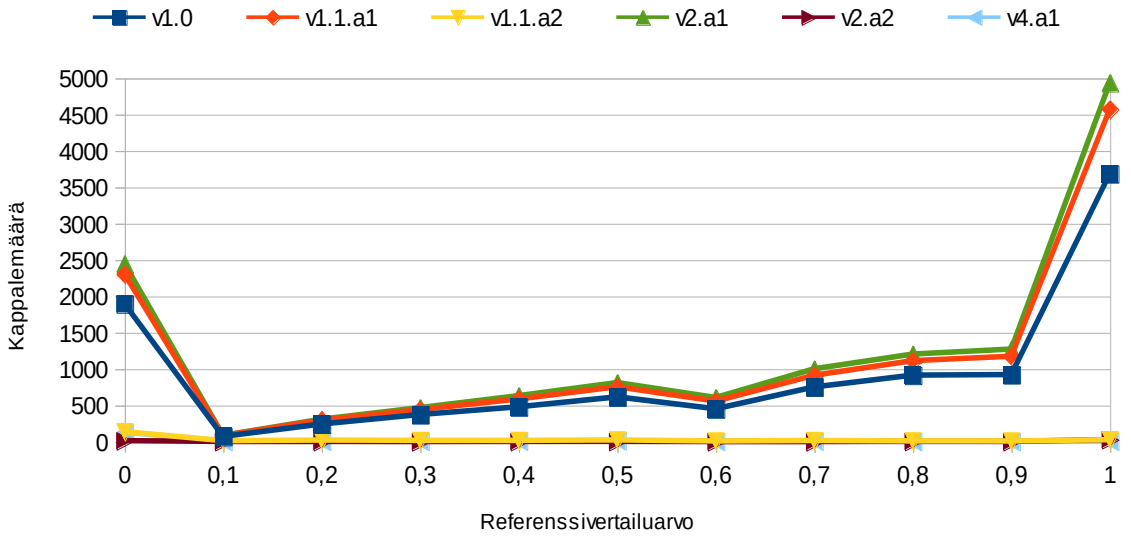
Kuva 9. Eksonitason laatuvertailuarvon jakauma tutkittavien annotaatioiden geeniennusteissa, joiden referenssivertailuarvo oli lähellä yhtä.

Jos taasen keskitymme vain laatuvertailuarvon mukaan lajittelemaan tuloksia huomaamme, että voimme jakaa tulokset niihin, joiden laatuvertailuarvo on alle 0,4 ja niihin, joissa se on yli 0,4 (kuva 10). Geeniennusteissa, joiden laatuvertailuarvo oli alle 0,4, näkyivät samat kolme datasettiä, kuin referenssivertailuarvon mukaan jaoteltuna, todella pienillä kappalemäärillä edustettuina (kuva 11). Tutkittaessa geeniennusteita,

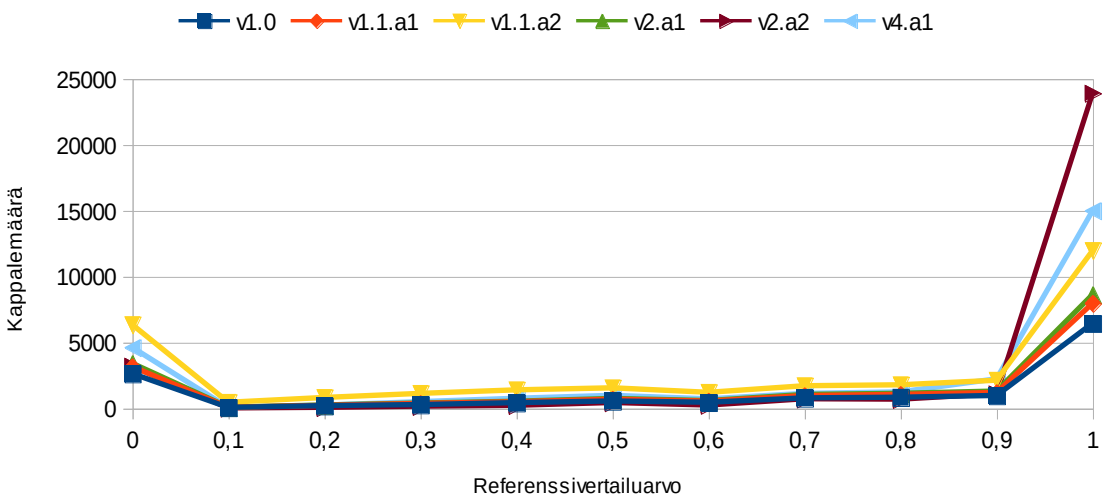
joiden laatuvertailuarvo oli yli 0,4, datasetillä v2.a2 oli eniten ennustettuja geenejä, jotka saivat referenssivertailuarvon lähellä yhtä (kuva 12).



Kuva 10. Laatuvertailuarvon jakauma kaikissa tutkittavissa annotaatioissa.



Kuva 11. Eksonitason referenssivertailuarvon jakauma tutkittavien annotaatioiden geeniennusteissa, joiden laatuvertailuarvo oli välillä 0 – 0,4.



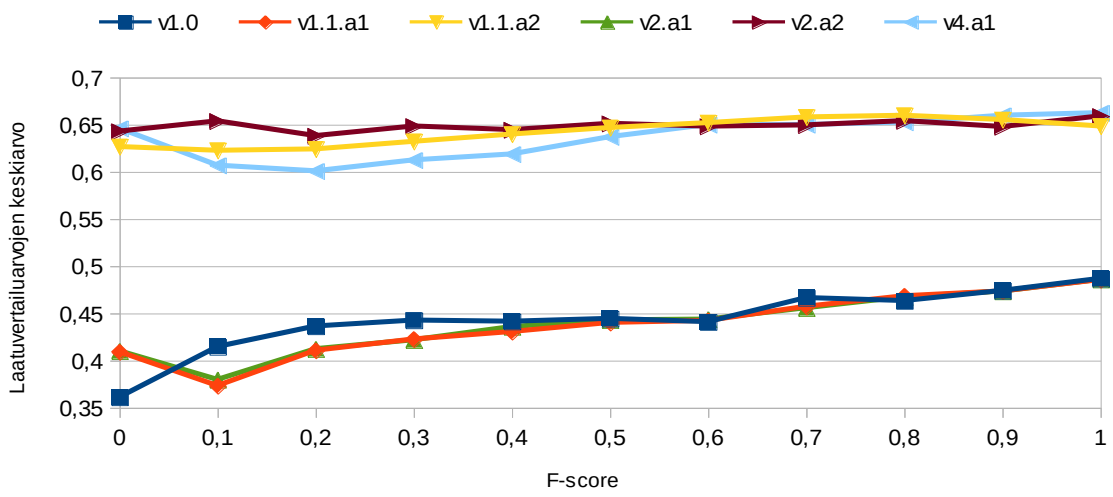
Kuva 12. Eksonitason referenssivertailuarvon jakauma tutkittavien annotaatioiden geeniennusteissa, joiden laatuvertailuarvo oli välillä 0,4 – 1.

Referenssivertailuarvon jakauma geeniennusteissa, jotka olivat saaneet alle 0,4 olevat laatuvertailuarvot (kuva 11), vastaa datasettien v1.0, v1.1.a1 ja v2.a1 osalta karkeasti referenssivertailuarvon jakaumaa kaikkien datasettien osalta (kuva 7). Jos jakauma olisi identtinen se viittaisi mahdollisuuteen, että referenssivertailuarvo ei korreloisi ollenkaan tutkittavien laatumittarien kanssa.

5.3 Nukleotiditaso

Koska nukleotiditason referenssiin vertauksessa käytetään vain herkkyyttä ja tarkkuutta käytämme f-scorea (5) kuvaamaan referenssiin vertauksen osuvuutta. Myös nukleotiditasolla verrattaessa f-scorea ja laatuvertailuarvoa keskenään jakautuivat datasetit samalla tavalla (kuva 13), kuin eksonitasolla vertailtaessa referenssivertailuarvoa ja laatuvertailuarvoa. Eksonitasolla havaittu mahdollinen korrelaatio referenssivertailuarvon ja laatuvertailuarvon välillä ei ollut enää nukleotiditasolla havaittavissa silmämääräisesti kaikissa dataseteissä. Annotaatiot v1.1a2, v2.a2 ja v4.a1 muodostivat jälleen kerran ryhmän, jonka laatuvertailu arvot olivat suurempia kuin toisella ryhmällä. Niiden F-score ei näyttänyt korreloivan laatuvertailuarvojen kanssa, vaikka korrelaatiokertoimista voidaan lukea korrelaation olemassaolosta (taulukko 5). Toisaalta datasettien v1.0, v1.1.a1 ja v2.a1 laatuvertailuarvojen keskiarvo muuttuu

nukleotiditasolla enemmän, kuin eksonitasolla, liikuttaessa f-scoren arvosta nolla yhteen.

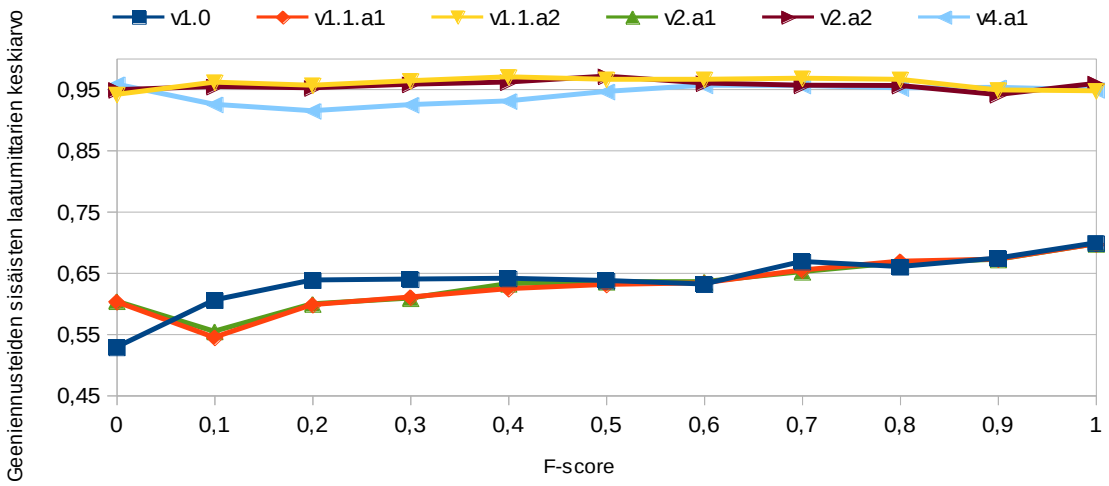


Kuva 13. Nukleotiditason f-score suhteessa kyseisen referessivertailuarvon saaneiden geeniennusteiden laatuvertailuarvojen keskiarvoon.

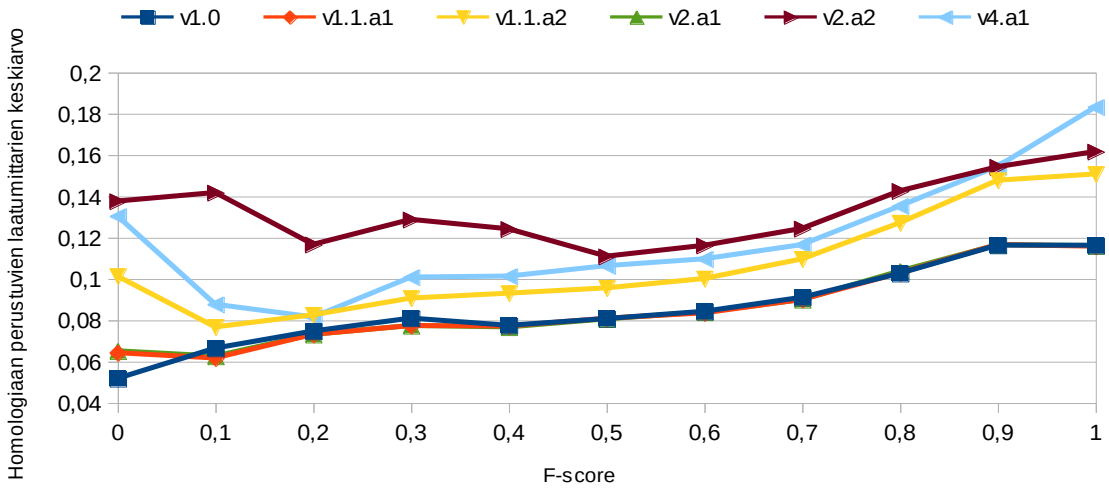
Taulukko 5. F-scoren, laatuvertailuarvon, geenien sisäisten laatumittarien ja proteiinien homologiaan perustuvien laatumittarien väliset korrelaatiokertoimet ja kaksisuuntaiset p-arvot.

		laatuvertailuarvo		sisäiset		homologiaan perustuvat	
annotaatio	otoskoko	korrelaatio	p-arvo	korrelaatio	p-arvo	korrelaatio	p-arvo
v4.a1	27733	0,1391	0,0000	0,0019	0,7578	0,3055	0,0000
v2.a2	30937	0,0678	0,0000	0,0191	0,0008	0,0996	0,0000
v2.a1	31957	0,1168	0,0000	0,0969	0,0000	0,2001	0,0000
v1.1.a2	30845	0,1072	0,0000	-0,0401	0,0000	0,2968	0,0000
v1.1.a1	29587	0,1200	0,0000	0,1004	0,0000	0,2008	0,0000
v1.0	24132	0,1251	0,0000	0,1041	0,0000	0,2123	0,0000

Suuremman laatuvertailuarvon ryhmän annotaatioiden geeniennusteiden sisäisten laatumittarien keskiarvot eivät muuttuneet f-scoren kasvaessa (kuva 14). Pienemmän laatuvertailuarvojen ryhmän datasetit ilmentävät pientä positiivista korrelaatiota f-scoren ja geeniennusteiden sisäisten laatumittarien keskiarvojen välillä. Proteiinien homologiaan perustuvissa laatumittarien keskiarvoissa liikuttaessa f-scoressa ylöspäin oli havaittavissa



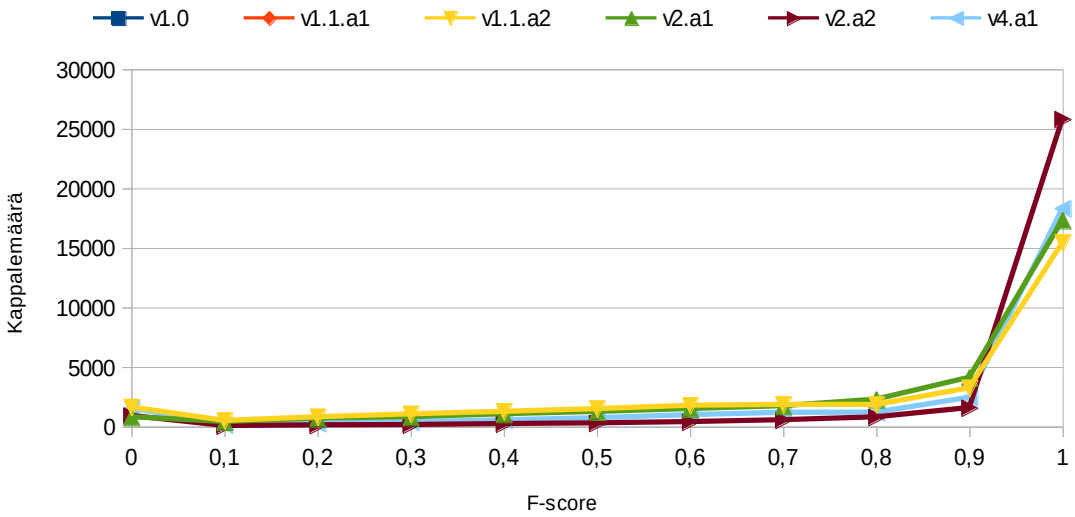
Kuva 14. Nukleotiditason referenssivertailuarvo suhteessa kyseisen referenssi-vertailuarvon saaneiden ennustettujen geenien geeniennusteiden sisäisten laatumittarien keskiarvoon.



Kuva 15. Nukleotiditason referenssivertailuarvo suhteessa kyseisen referenssi-vertailuarvon saaneiden geeniennusteiden proteiinien homologiaan perustuvien laatumittarien keskiarvoon.

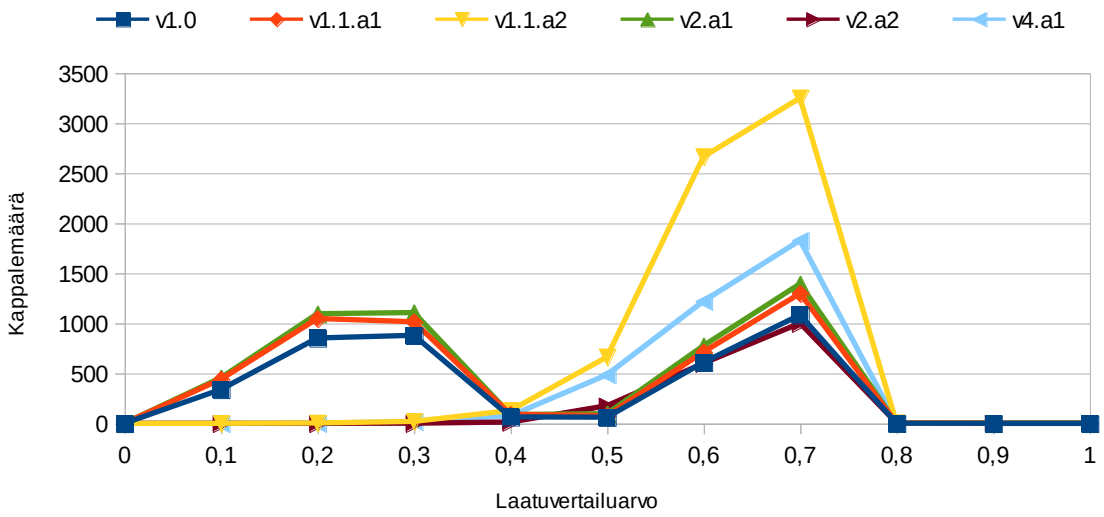
selkeä ylöspäin suuntautuva trendi (kuva 15), vaikka erot eri f-scoren saaneiden laatumittarien keskiarvojen välillä olivat todella pieniä.

Jos järjestämme datasetit f-scoren mukaan huomaamme, että suurin osa jokaisen datasetin geeniennusteista oli saanut lähellä yhtä olevan f-score arvon (kuva 16). Näin ollen tutkimme laatumittarien keskiarvoja kolmessa osassa, ensin niiden geeniennusteiden osalta, joiden f-score oli 0 – 0,5, sitten 0,6 – 0,9 ja lopulta niiden osalta, joiden f-score oli lähellä yhtä.

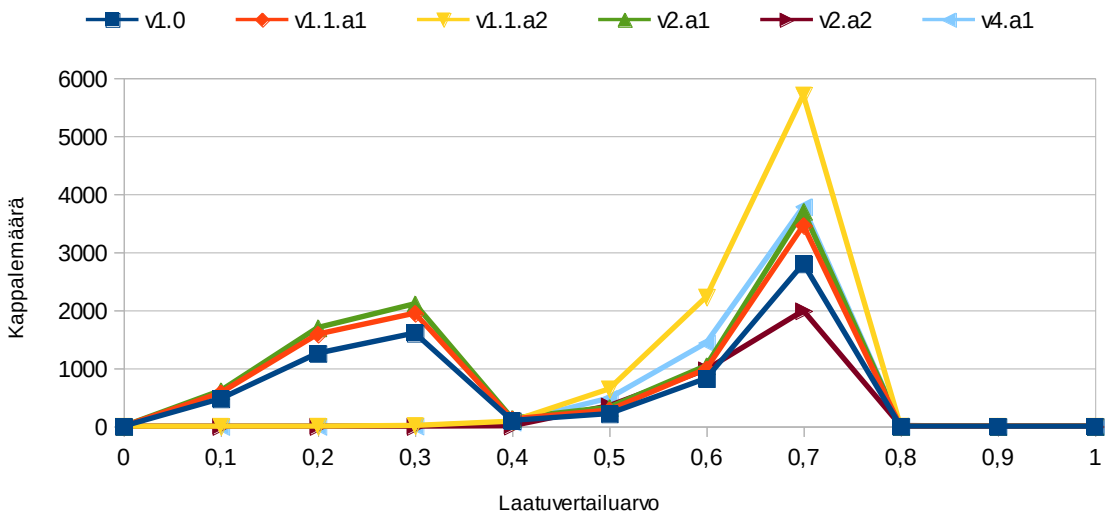


Kuva 16. Nukleotiditason f-scoren jakauma tutkittavissa annotaatioissa.

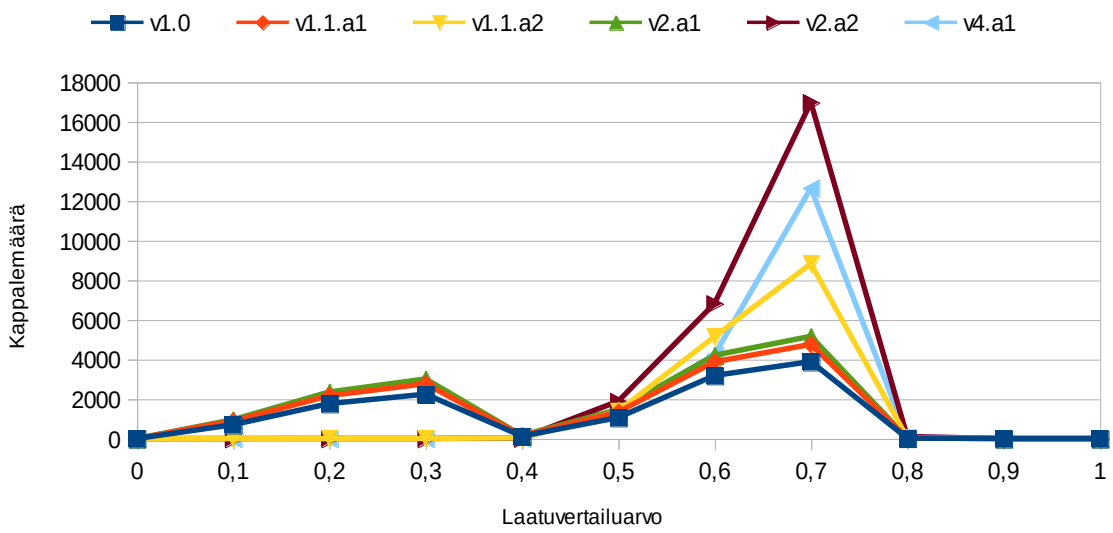
Geeniennusteissa, joiden f-score oli välillä 0 – 0,5, datasetit v1.0, v1.1.a1 ja v2.a2 jakautuivat kahteen selkeästi toisistaan erossa oleviin keskittymiin, jotka sisälsivät karkeasti saman verran geeniennusteita (kuva 17). Geeniennusteissa, joiden f-score oli välillä 0,6 – 0,9, kappalemäärällisesti eniten suuren laatumittarien keskiarvon omaavia geeniennusteita oli datasetissä v1.1.a1 (kuva 18). Tarkasteltaessa geeniennusteita, jotka saivat f-scoren lähellä yhtä, datasetit v1.1.a2, v2.a2 ja v4.a1 sisälsivät eniten suuren laatumittarien keskiarvon saaneita geeniennusteita (kuva 19). Nämä vertailut avartavat datasettien v1.0, v1.1.a1 ja v2.a1 laatuvertailuarvon nousua liikuttaessa f-scoren arvosta nolla yhteen. Annotaatioiden laatuvertailuarvon jakauma voidaan jakaa pienemmän laatuvertailuarvon ryhmään ja suuremman laatuvertailuarvon ryhmään. F-scoren noustessa pienemmän laatuvertailuarvon ryhmä pienenee suhteessa toiseen ryhmään.



Kuva 17. Nukleotiditason laatuvertailuarvon jakauma tutkittavien annotaatioiden geeniennusteissa, joiden f-score oli välillä 0 – 0,5.

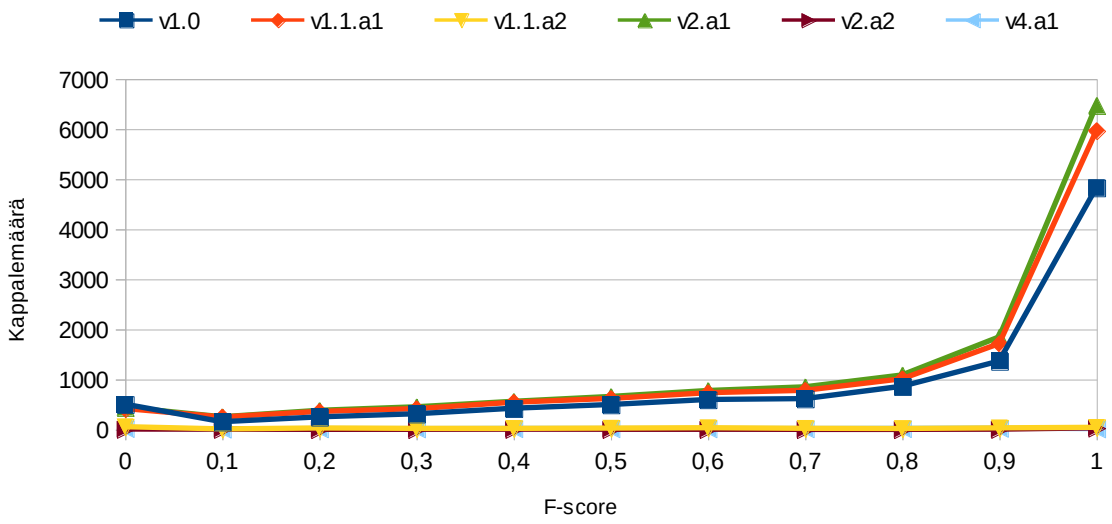


Kuva 18. Nukleotiditason laatuvertailuarvon jakauma tutkittavien annotaatioiden geeniennusteissa, joiden f-score oli välillä 0,6 – 0,9.

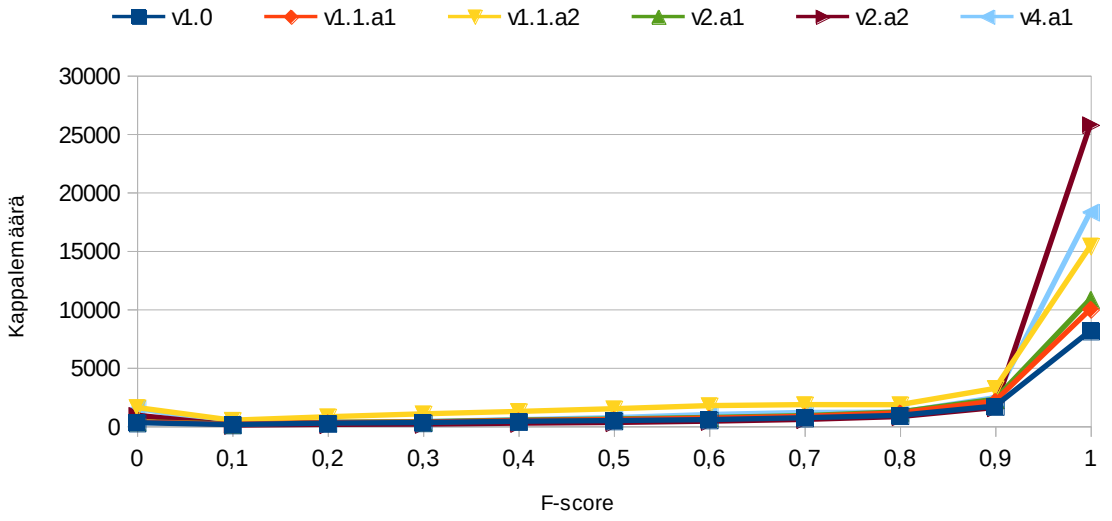


Kuva 19. Nukleotiditason laatuvertailuarvon jakauma tutkittavien annotaatioiden geeniennusteissa, joiden f-score oli lähellä yhtä.

Koska eksonitasosolla ja nukleotiditasolla tutkimme samoja geenejä, niiden laatumittarien keskiarvon jakauma on sama (kuva 6) ja voimme jakaa tulokset ennusteisiin, joiden laatumittarien keskiarvo oli alle 0,4 ja niihin, joissa se oli yli 0,4. Tarkasteltaessa geeniennusteita, joiden laatumittarien keskiarvo oli alle 0,4, datasettien v1.0, v1.1.a1 ja v2.a1 f-scoren jakaumassa on paljon geeniennusteita, joiden f-score on lähellä yhtä (kuva 20). Geeniennusteissa, joiden laatuvertailuarvo oli yli 0,4, f-scoren jakaumassa jokaisella datasetillä oli eniten geeniennusteita lähellä f-scoren arvoa yksi (kuva 21). Ne myös sisälsivät hiukan enemmän matalan f-scoren saaneita geeniennusteita, jotka myös saivat laatuvertailuarvon välillä 0 – 0,4, kuin matalan f-scoren saaneita geeniennusteita, jotka saivat laatuvertailuarvon välillä 0,5 – 1



Kuva 20. Nukleotiditason f-scoren jakauma tutkittavien annotaatioiden geeniennusteissa, joiden laatuvertailuarvo oli välillä 0 – 0,4.



Kuva 21. Nukleotiditason f-scoren jakauma tutkittavien annotaatioiden geeniennusteissa, joiden laatuvertailuarvo oli välillä 0,4 – 1.

5.4 Yhteenvetona

Suurimmassa osassa tuloksia havaittava datasettien jakautuminen kahteen ryhmään johtui siitä, että hieman alle puolet pienemmän numeraalisen arvon ryhmän annotaatioiden geeniennusteista sai hyvin pienen arvon yhdestä tai useammasta geeniennusteiden sisäisestä laatumittarista. Luultavimpia vaihtoehtoja olivat laatumittarit, jotka pystyivät saamaan arvon vain arvon yksi tai nolla. Tällaisia laatumittareita olivat teoreettisen proteiinin alkaminen metioniini aminohapolla ja lopetus kodonin löytyminen.

Nukleotiditasolla huomattu hyvin korkea geeniennusteiden sisäisten laatumittarien keskiarvo, joka ei vaikuttanut olevan silmämääräisesti linkittynyt f-scoreen, tarkoittaa, että suurin osa geeniennusteista sai melkein täydellisen arvon laatumittareista ja jäljelle jääneet pienemmän arvon geeniennusteet olivat jakautuneet f-scoressa tasaisesti. Geeniennusteiden sisäisten laatumittarien korrelaatiokertoimet nukleotiditasolla (taulukko 5) viittaavat siihen, että kyseisten laatumittarien korrelaatio heikkenee tutkittavien geeniennusteiden laadun parantuessa. Tämä on jopa hiukan itsestään selvää, koska geenien ennustusohjelmat rakennetaan etsimään geenejä, jotka täyttävät geenien perus rakenteen. On jopa hämmentävää, kuinka paljon virheellisiä geeniennustuksia pääsee ohjelmista läpi.

6 Johtopäätökset

Tutkimuksen tuloksia tulkittaessa täytyy ottaa huomioon, että valittu referenssi annotaatio ei ole täydellinen mallinnus mansikan genomista. Se on vain uusin ja hyvin luultavasti tarkin versio yrityksestä kuvata pientä osaa todellisuuden toiminnasta. Aikaisemmat tuotetut annotaatiot, joita on käytetty tässä tutkimuksessa aineistona eivät eroa pohjimmiltaan datasetistä, joka valittiin referenssiksi. Näin ollen tutkimus olisi ehkä ollut järkevämpi suorittaa valiten tutkimusaineisto siten, että referenssiksi voidaan valita todella paljon laadunvalvontaa kokenut annotaatio. Koska osa vertailuista ja analyyseistä tehtiin itse tehdyillä ohjelmilla, niissä voi olla virheitä, jotka voivat vaikuttaa tuloksiin, mutta usein kyseiset virheet olivat helppoja löytää.

Tutkimuksen tulokset puoltavat ajatusta, että valitut geeniennusteiden sisäiset laatumittarit eivät toimi järkevänä geeniennusteiden laadun mittarina. Niiden korrelaatio vaikuttaisi tippuvan mitä parempilaatuisia geeniennusteita niillä yritetään arvioida. Toisaalta ehkä niitä olisi mahdollista käyttää huonolaatuisten annotaatioiden automaattiseen etsintään. Ennalta odottamattomin tulos olivat korrelaatiokertoimien yllättävän hyvät tulokset. Ne eivät ole suuria, mutta koska otoskoot olivat niin suuria niiden kaksisuuntaiset p-arvot ovat todella pieniä. Pienikin signaali on mahdollista havaita kohinan seasta, jos otantakoko on tarpeeksi suuri. Proteiinien homologiaan perustuvat laatumittarit vaikuttivat lupaavilta laadun arvioinnissa.

Laskettaessa laatuvertailuarvoja kaikilla käytetyillä laatumittareilla oli yhtä suuri vaikutus sen suuruuteen. Mahdollisissa jatkotutkimuksissa voidaan parantaa laatuvertailuarvon toimintaa antamalla laatumittareille painotuksia ja lisäämällä uusia laatumittareita. On tietenkin mahdollista, että on olemassa joukko geenien rakenteeseen perustuvia laatumittareita, joilla voidaan tehdä korkealaatuisten geeniennusteiden laadunvalvontaa.

7 Lähteet

Burset, M. & Guigo, R. 1996. Evaluation of Gene Structure Prediction Programs. GENOMICS 34, 353–367.

Campbell, M. S., Law, M. Y., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S. H., Childs, K. L., Sun, Y., Jiang, N. & Yandell, M. 2014. MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. PLANT PHYSIOLOGY 164: 513–524.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S. & Yandell, M. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18: 188–196.

Darwish, O., Shahan, R., Liu, Z., Slovin, J. P. & Alkharouf, N. W. 2015 Re-annotation of the woodland strawberry (Fragaria vesca) genome. BMC genomics 16(1): 29.

Dragan, M. A., Moghul, I., Priyam, A., Bustos, C. & Wurm, Y. 2016. BIOINFORMATICS 32(10): 1559–1561.

Edger, P. P., VanBuren, R., Colle, M., Poorten, T. J., Wai, C. M., Niederhuth, C. E., Alger, E. I., Ou, S., Acharya, C. B., Wang, J., Callow, P., McKain, M. R., Shi, J., Collier, C., Xiong, Z., Mower, J. P., Slovin, J. P., Hytönen, T., Jiang, N., Childs, K. L. & Knapp, S. J. 2017. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (Fragaria vesca) with chromosome-scale contiguity. GIGASCIENCE 7(2).

Hoff, K. F. & Stanke, M. 2013. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Research 41(w1): w123-w128.

Holt, C. & Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC BIOINFORMATICS 12: 491.

Jung, S,. Lee, T., Cheng, C. H., Buble, K., Zheng, P., Yu, J., Humann, J., Ficklin, S. P., Gasic, K., Scott, K., Frank, M., Ru, S., Hough, H., Evans, K., Peace, C., Olmstead, M., DeVetter, L. W., McFerson, J., Coe, M., Wegrzyn, J. L., Staton, M. E., Abbott, A. G. & Main, D. 2019. 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. NUCLEIC ACIDS RESEARCH 47: D1137-D1145.

Korf, I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5: 59.

Korf, I. 2015. http://korflab.ucdavis.edu/Datasets/cegma/. viitattu 31.10.2020.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simao, F. A. & Zdobnov E. M. 2018. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. NUCLEIC ACIDS RESEARCH 47: D807-D811.

Li, Y. P., Pi, M. T., Gao, Q., Liu, Z. C. & Kang, C. Y. 2019. Updated annotation of the wild strawberry Fragaria vesca V4 genome. HORTICULTURE RESEARCH 6: 61.

Li, Y., Wei, W., Feng, J., Luo, H., Pi, M., Liu, Z. & Kang, C. 2018. Genome re-annotation of the wild strawberry Fragaria vesca using extensive Illumina- and SMRT-based RNA-seq datasets. DNA RESEARCH 25: 61-70.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. NUCLEIC ACIDS RESEARCH 33(20): 6494-6506.

Löytynoja, A. 2018. Kurssimateriaali. http://loytynojalab.biocenter.helsinki.fi/data/evogeno/PracticalGenomicalignment.pdf. viitattu 31.10.2020.

Parra, G., Bradnam, K. & Korf, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9): 1061-1067.

Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. Nucleic Acids Research 37: 289-297.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J. M., Rees, D. J. G., Williams, K. P., Holt, S. H., Rojas, J. J. R., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Viola, R., Ashman, T. L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant, D. W., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Girona, E. L., Zdepski, A., Wang, W. Q., Kerstetter, R. A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E. & Folta, K. M. 2011a. The genome of woodland strawberry (Fragaria vesca). NATURE GENETICS 43(2): 109-116.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J. M., Rees, D. J. G., Williams, K. P., Holt, S. H., Rojas, J. J. R., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Viola, R., Ashman, T. L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant, D. W., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Girona, E. L., Zdepski, A., Wang, W. Q., Kerstetter, R. A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E. & Folta, K. M. 2011b.

https://www.rosaceae.org/species/fragaria/fragaria_vesca/genome_v1.1. viitattu 31.10.2020.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. BIOINFORMATICS 31(19): 3210-3212.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. & Morgenstern, B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. NUCLEIC ACIDS RESEARCH 34: W435-W439.

Stanke, M. & Waack, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. BIOINFORMATICS 19: ii215–ii225.

Tennessen, J. A., Govindarajulu, R., Ashman, T. L. & Liston, A. 2014. Evolutionary Origins and Dynamics of Octoploid Strawberry Subgenomes Revealed by Dense Targeted Capture Linkage Maps. GENOME BIOLOGY AND EVOLUTION 6(12): 3295-3313.